MULTI-RESOLUTION LC-MS IMAGES ALIGNMENT USING DYNAMIC TIME WARPING AND KULLBACK-LEIBLER DISTANCE

William K. H. Wu [1], *Albert C. S. Chung* [1] *and Henry H. N. Lam* [2]

Lo Kwee-Seong Medical Image Analysis Laboratory, Dept. of CSE [1]
Dept. of CBME [2], The Hong Kong University of Science and Technology, Hong Kong

ABSTRACT

Liquid chromatography mass spectrometry (LC-MS) is widely used in comparing proteomes for disease biomarker discovery. An LC-MS experiment produces a 2-D image, where the mass-to-charge ratio and the chromatographic retention time are the coordinates, and the signal intensities represent the abundance of detected peptides. However, there is always a non-linear retention time difference across replicate LC-MS images due to machine drift, such that synchronization of LC-MS images must be performed prior to any further analysis.

In this paper, we propose a multi-resolution image alignment scheme to synchronize LC-MS images. Dynamic Time Warping (DTW) is used to reconcile the time differences among images and Kullback-Leibler distance (KLD) is used as a local distance measure. Our proposed scheme has been validated using two real data sets, and promising results have been obtained.

Index Terms— multi-resolution, LC-MS, DTW, KLD

1. INTRODUCTION

Detection and identification of chemicals in a complex mixture is crucial in many chemical analyses, such as the identification of peptides in the study of proteomics. Liquid chromatography mass spectrometry (LC-MS) is one of the most widely used analytical chemistry technique for determining the elemental composition of chemical compounds. LC-MS combines the physical separation capabilities of liquid chromatography with the mass analysis capabilities of mass spectrometry to provide a two-dimensional approach to compare chemical mixtures by both the mass-to charge ratio and LC retention time. The resulting 2D LC-MS image can be viewed as a profile of all chemicals in a mixture. One major problem in analyzing LC-MS images is the problem of time shifting in the retention time dimension. Even for the same chemical mixtures, different LC-MS images are produced across different runs of LC-MS experiments because the retention time dimension of LC-MS images varies non-linearly due to machine drift, which means that the produced LC-MS images are stretched and shrunk locally compared to each other. Without accurate time alignment between two LC-MS images, non-linear retention time shift of two images will lead to incorrect statistical comparisons. As a result, synchronization of LC-MS images must be done prior to performing any further analysis.

In this paper, a multi-resolution LC-MS image alignment scheme using Dynamic Time Warping (DTW) alogrithm [1] and Kullback-Leibler distance (KLD) [2] is proposed. DTW is a popular dynamic programming algorithm for reconciling the time difference between two sequences with different lengths which minimize the effects of local time shifting by allowing the sequences locally translated, compressed and expanded. Besides, KLD is employed in combined with DTW as a local distance measure for alignment. Although DTW can yield an optimal solution, it is not cost effective to accomplish this by searching the whole solution space, and thus several constraints are proposed to confine the search space [3, 4]. However, this causes a problem when the optimal solution lies outside the search space. To overcome this problem, multi-resolution images are down-scaled from the original LC-MS image, and the optimal solution is found iteratively. Comparisons of time alignment quality with an existing time alignment algorithm have been performed. We have also validated our alignment scheme using two real data sets, and promising results have been found.

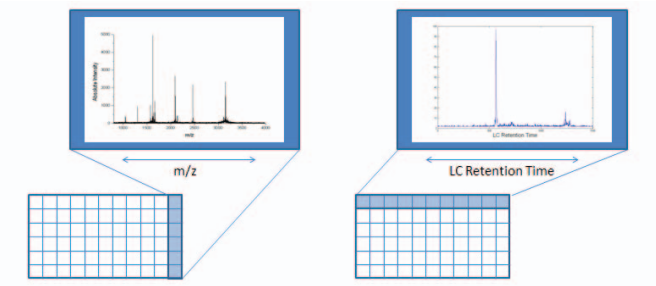
2. LC-MS IMAGE ALIGNMENT

In this section, a brief description of the structure of LC-MS images is presented. Then, the nature difference between LC-MS image alignment and traditional image registration is also examined. Finally, some approaches developed to align LC-MS images are described.

2.1. Image Structure

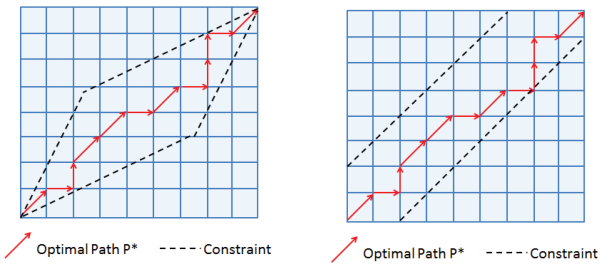
A LC-MS image can be viewed as a sequence of mass spectra produced at different scan time during a chemical experiment where the LC retention time and mass-to-charge ratio (m/z) are the x-axis and the y-axis of the image respectively.

A LC-MS image can provide a two-dimensional approach to analyze a chemical mixture by both the mass spectra and



(a) Mass Spectrum (b) Liquid Chromatogram

Fig. 1: LC-MS Image viewing along x and y-axis.



(a) Itakura global constraint (b) Sakoe-Chiba band constraint

Fig. 2: Typical constraint.

liquid chromatograms. The mass spectra are represented by column vectors at different retention time as shown in Figure 1a and the liquid chromatograms are represented by row vectors at different mass traces as shown in Figure 1b.

2.2. Problem Nature

Aligning LC-MS images is different from ordinary image registration problems because traditional image registration methods do not fit the nature of the problem. Rigid image registration methods cannot be applied because the target LC-MS image is generally stretched and shrunk locally compared to the source image. Although non-rigid image registration methods can determine the correspondence in two LC-MS images by some localized stretching of the images, those methods have to enforce the constraint that LC-MS images are only stretched and shrunk along one dimension (i.e. the retention time dimension). More importantly, it is likely to get stuck into a local minimum because the objective function should not be simply convex, yet with many local minima. Since it is known that LC-MS images are aligned column by column, instead of solving this alignment problem by a 2D image registration method, we align the images by the Dynamic Time Warping (DTW) algorithm.

2.3. Related Work

A few approaches have been introduced to align LC-MS images using DTW. For example, the simplest method is to use one-dimensional profiles such as the Total Ion Chromatogram (TIC) or Base Peak Chromatogram (BPC) to represent the entire 2D LC-MS images, which works only if the images have a relatively simple structure. This method often leads to misalignment of LC-MS images produced from complex mixtures because compounds with different m/z values but eluting at the same retention time are not considered separately by using only one-dimensional information. Some approaches treat the information of different mass traces separately in the local distance used by DTW to ensure that the intensity information of compounds with different masses is not mixed. For example, Christin et al. [5] use a Component Detection Algorithm (CODA) [6] to select high quality mass traces from the entire image prior to performing alignment.

3. METHODOLOGY

In this section, some background information about the theory of Dynamic Time Warping (DTW) is presented. Moreover, the distance measure employed in this paper, the Kullback-Leibler distance (KLD), is also described. Finally, our proposed multi-resolution LC-MS image alignment scheme is explained in detail.

3.1. Dynamic Time Warping

Dynamic Time Warping (DTW) is a widely used algorithm for finding an optimal match between two sequences with different lengths by non-linearly warping the sequences such that similar objects are aligned and the overall distance between them is minimized. Consider two sequences:

$$X_n = x_1 x_2 ... x_i ... x_n$$
$$Y_m = y_1 y_2 ... y_i ... y_m$$

To align the sequences, DTW constructs an n-by-m matrix called the warping matrix where the $(i,j)th$ element of the matrix is the minimum accumulated distance of the optimal warping $D(i,j)$ for the subsequences X_i and Y_j. The minimum accumulated distance of the optimal warping for the whole sequences X_n and Y_m, namely $D(n,m)$, can be found by solving the following optimization problem:

$$D(i,j) = \min \left\{ \begin{array}{l} D(i-1,j) + d(i,j) \\ D(i-1,j-1) + 2 \cdot d(i,j) \\ D(i,j-1) + d(i,j) \end{array} \right\}$$

where $d(i,j)$ is the local distance between the objects x_i and y_j.

By storing the predecessor of each element in the warping matrix, an optimal warping path P^*, consists of a sequence

of order pairs, can be reconstructed by backtracking (see Fig. 2.). The optimal path indicates how the sequences should be aligned, for example, an order pair (i, j) in the warping path means that objects x_i and y_j should be aligned together.

In general, some constraints should be applied to limit the search space in order to reduce the running time. Figure 2a shows the Itakura global constraint [3] and Figure 2b shows the band constraint proposed by Sakoe and Chiba [4]. The Sakoe-Chiba band constraint forbids the optimal path to deviate $\pm M$ points from the linear path starting at point $(1, 1)$. To include the destination point (n, m) in the search space, M has to be equal to or greater than the absolute difference between n and m, that is,

$$M \geq |n - m|$$

3.2. Kullback-Leibler Distance

As stated above, a local distance measure, which measures the distance between objects of the sequences, has to be defined in order to apply DTW. Recall that the LC-MS images are aligned column by column, so each element belonging to the same column should be moved as a whole. As a result, "objects" represent columns of a LC-MS image, that is, the mass spectra scanned at different retention time (see Fig. 1a.). In order words, a distance measure suited to evaluate the difference between two mass spectra is required.

In this study, Kullback-Leibler distance (KLD), which is a widely used measure in the field of information theory, is employed as the local distance measure. KLD is a non-symmetric measure originally used to evaluate the difference between two probability distributions, which in turn can measure the difference between two mass spectra:

$$D(P_{x_i} \parallel P_{y_j}) = \sum_m P_{x_i}(m) ln \frac{P_{x_i(m)}}{P_{y_j(m)}}$$

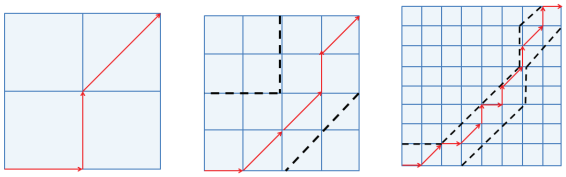
However, KLD is only defined if $P_{y_j}(m) > 0$ for any mass traces m such that $P_{x_i}(m) > 0$, otherwise $D(P_{x_i} \parallel P_{y_j})$ will become infinite. To fulfill this requirement, an insignificant intensity value ε is added to the whole image as background intensity. Moreover, the standard KLD is a non-symmetric distance measure, meaning that it is not a true metric. We therefore employ the symmetric version of KLD as follows:

$$D(P^*_{x_i} \parallel P^*_{y_j}) = \sum_m \left(\left(P^*_{x_i}(m) - P^*_{y_j}(m)\right) ln \frac{P^*_{x_i(m)}}{P^*_{y_j(m)}} \right)$$

where $P^*(m)$ is the probability after adding the background intensity ε. When two mass spectra are similar to each other, the value of $D(P^*_{x_i} \parallel P^*_{y_j})$ is expected to be small.

3.3. Multi-Resolution LC-MS Image Alignment

As shown in Figure 2, different constraints are proposed to reduce the search space of the optimal path P^*. However, the actual optimal path P^* may lie outside the search area, which will result in a failure alignment.



(a) s = 4 (b) s = 2 (c) s = 1

Fig. 3: Multi-Resolution LC-MS Image Alignment.

As a result, a multi-resolution LC-MS image alignment scheme is proposed. Firstly, the original image I is down-scaled by a factor of s, constructing several down-scaled images at different levels, I_s, such that $I_s = I$ when $s = 1$. In this study, three resolution levels are implemented, i.e. $s = 1, 2, 4$. Notably, the original image is not only down-scaled along the retention time dimension, but also the m/z dimension, resulting in "shorter" mass spectra at the lower resolution level. Secondly, DTW is performed iteratively starting from the lowest resolution level to the highest resolution level (i.e. $s = 1$), and the warping path found at the previous resolution level $2s$ is used to confine the search space S_s of the current resolution level s as follows:

$$S_s = P_{2s} \pm M$$

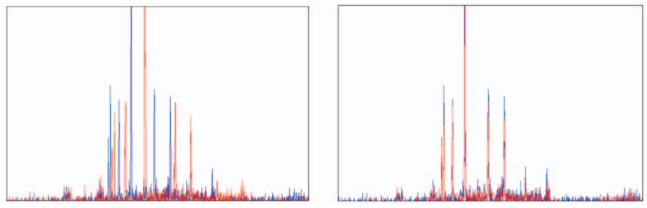
where $M = |n - m| * s$. At the lowest resolution level, P_{2s} is equal to the linear path starting at point $(1, 1)$, and the constraint becomes the Sakoe-Chiba band constraint. Figure 3 demonstrates the iterative process of finding the optimal path. By searching a warping path in a lower resolution warping matrix, the area of the search space is actually increased, resulting in a higher chance of including the optimal warping path in the next resolution level.

4. EXPERIMENTAL VALIDATION

To validate our proposed LC-MS image alignment scheme, several experimental results with two real-world data sets of LC-MS images are presented. Apart from statistical results, graphical illustration is also provided to show how LC-MS images are aligned. Comparison of time alignment quality between our proposed scheme and DTW-CODA [5] have been also performed.

4.1. Set-up

The experiment was run on a server with Intel Core i7 @ 2.67 GHz and 12GB of RAM. Two real-world data sets used in the experiments are `QS` and `MOUSE`. `QS` is a real-world data set consists of 3 pairs of LC-MS images with high mass accuracy produced from the tryptic digests of human blood serum



(a) Before Alignment (b) After Alignment

Fig. 4: Comparsion of mass traces before and after alignment.

spiked with six standard nonhuman proteins at different concentrations. MOUSE is a real-world data set consists of 17 LC-MS images with low mass accuracy of mouse blood samples.

To validate the alignment accuracy, we randomly generated 100 synthetic LC-MS images for each real LC-MS image. Compared to its corresponding real LC-MS image, those synthetic images are shifted non-linearly along the retention time dimension. In total, 2300 synthetic LC-MS images were generated and each of them was aligned with the corresponding real LC-MS image. Since the synthetic images are artificially generated, the optimal warping path can be obtained as a benchmark. For each alignment, we calculated the root-mean-squared deviation (RMSD) between the experimental warping path and the optimal warping path. Moreover, we defined a criteria that, for each alignment, RMSD has to be less than 1 scan in order to be claimed as successful.

4.2. Statistical Results

Table 1 shows the average success rate of alignment. Experimental results showed that our proposed method perform better than DTW-CODA for both data sets. For QS and MOUSE, our method (3 resolution levels) outperformed DTW-CODA for about 30% and 25% in the average success rate respectively. It is also observed that some unsuccessful alignment of DTW-CODA was caused by an optimal path lying outside the search space, while our method could obtain a successful alignment for those cases.

4.3. Graphical Illustration

Apart from numerical verification, Figure 4 provides a graphical illustration to demonstrate how real LC-MS images are aligned by our proposed method. We selected the same mass trace from a pair of LC-MS images in QS and the cross-section of the images before and after alignment are showed respectively. As shown in the figures, the originally unaligned peptide peaks are matched to the same retention time after performing the alignment, and hence the LC-MS images are correctly aligned.

Table 1: Success rate Comparsion

Success rate (%)	QS	Mouse
Our method (3 resolution levels)	96.50	97.88
Our method (2 resolution levels)	87.33	88.82
DTW-CODA	66.17	72.47

5. CONCLUSION

In this paper, we propose a multi-resolution image alignment scheme to synchronize LC-MS images. By down-scaling the LC-MS images at different resolution levels, the Dynamic Time Warping (DTW) algorithm is used in combined with the Kullback-Leibler distance (KLD) for iterative alignment.

Our proposed scheme has been validated using two real data sets and promising results have been demonstrated. Our method has also been compared with an existing alignment algorithm DTW-CODA, experimental results showed that our method performs better than DTW-CODA in alignment accuracy. For future work, we will consider the possibility of various local distance measures used in combined with DTW.

6. REFERENCES

[1] Athanassios Kassidas, John F. MacGregor, and Paul A. Taylor, "Synchronization of batch trajectories using dynamic time warping," *AIChE Journal*, vol. 44, no. 4, pp. 864–875, 1998.

[2] Thomas M. Cover and Joy A. Thomas, *Frontmatter and Index*, pp. i–xxiii, John Wiley & Sons, Inc., 2001.

[3] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 23, no. 1, pp. 67 – 72, feb 1975.

[4] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 1, pp. 43 – 49, feb 1978.

[5] Christin Christin, Huub C. J. Hoefsloot, Age K. Smilde, Frank Suits, Rainer Bischoff, and Peter L. Horvatovich, "Time alignment algorithms based on selected mass traces for complex lc-ms data," *Journal of Proteome Research*, vol. 9, no. 3, pp. 1483–1495, 2010, PMID: 20070124.

[6] Willem Windig, J. Martin Phalp, and Alan W. Payne, "A noise and background reduction method for component detection in liquid chromatography/mass spectrometry," *Analytical Chemistry*, vol. 68, no. 20, pp. 3602–3606, 1996.